Our Ref. No.042390.P13129
Express Mail No.: EL 802886868 US

UNITED STATES PATENT APPLICATION

FOR

MULTI-LEVEL RING PEER-TO-PEER NETWORK STRUCTURE FOR PEER AND OBJECT DISCOVERY

Inventors:

Gururaj Nagendra
Shaofeng Yu

Prepared by:

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP
12400 Wilshire Blvd., 7th Floor
Los Angeles, CA 90025-1026
(714) 557-3800

MULTI-LEVEL RING PEER-TO-PEER NETWORK STRUCTURE FOR PEER AND OBJECT DISCOVERY

BACKGROUND

1. Field of the Invention

[001] This invention relates to networks, and in particular, the invention relates to peer-to-peer communication.

2. Description of Related Art

[002] Peer-to-Peer (P2P) communication has recently become popular. P2P communication allows a computer or a network device to communicate or locate another computer or network device across a network. Currently, computers and other Internet protocol (IP)-based devices can be discovered using the existing Domain Name System (DNS) or using P2P mechanisms like Napster or Gnutella.

[003] These techniques for P2P communication have a number of disadvantages. Napster and DNS are server-based name services. More particularly, the DNS is a distributed Internet directory service. DNS is used mostly to translate between domain names and IP addresses, and to control Internet message delivery. In server-based name services such as DNS and Napster-like P2P systems, the name servers must maintain a large list of computers or network devices that they support. For a large number of names, DNS needs complex software and processing power for name resolutions. It also requires heavy administration and maintenance. Moreover, DNS-like techniques may contain state IP addresses because of the typically high Time-To-Live (TTL) validity in resource records.

[004] In distributed systems like Gnutella, a message is required to traverse through the network one peer at a time and branches out until the result is found in a peer. This traversing generates a lot of network traffic, especially when the amount of messages is large. Since Gnutella employs a flat model, it may not scale well with the network. It also does not maintain any index or cache of other computers and may cause performance

degradation when the network grows. In addition, it is difficult and inconvenient to form and name groups of computers in these systems.

[005] Another technique uses bootstrap servers for its operation (e.g., Legion). These bootstrap servers run agents or managers that manage various requests on the P2P network or agents. If a peer needs to discover another peer, it may find it in the directory structure provided by the managers. One disadvantage of this technique is that the bootstrap server is a single point of failure. A failure of the bootstrap server may render the service non-operational.

[006] Therefore, there is a need to have an efficient technique to provide P2P communication.

BRIEF DESCRIPTION OF THE DRAWINGS

[007] The features and advantages of the present invention will become apparent from the following detailed description of the present invention in which:

[008] Figure 1 is an exemplary diagram illustrating a system 100 in which one embodiment of the invention can be practiced;

[009] Figure 2 is an exemplary diagram illustrating a peer shown in Figure 1 according to one embodiment of the invention;

[0010] Figure 3 is an exemplary diagram illustrating a P2P subsystem shown in Figure 2 according to one embodiment of the invention; and

[0011] Figure 4 is an exemplary flowchart illustrating a process for P2P communication using a hierarchical ring structure according to one embodiment of the invention.

DESCRIPTION OF THE INVENTION

[0012] The invention is a technique to provide Peer-to-Peer (P2P) communication among peers. In one embodiment of the invention, a P2P subsystem includes a cache of a current peer and a peer locator. The current peer is in a current ring at a current level. The cache stores information of ring peers within the current ring. The current ring is part of a hierarchical ring structure of P2P nodes. The hierarchical ring structure has at least one of a lower level and an upper level. The peer locator locates a target peer in the cache in response to a request. If the target peer is not found in the cache, the current peer forwards the request to either the other peers in the current ring at the current level or a peer at an upper level.

[0013] The technique in the present invention achieves at least the following advantages: (1) A server software is not needed, (2) Insignificant or minimum administration and maintenance, (3) No zone file containing all the peers is needed, (4) No stale Internet Protocol (IP) addresses are kept for a long time, (5) Less network traffic than the distributed technique, and (6) High scalability.

[0014] In the following description, for purposes of explanation, numerous details are set forth in order to provide a thorough understanding of the present invention. However, it will be apparent to one skilled in the art that these specific details are not required in order to practice the present invention. In other instances, well-known structures are shown in block diagram form in order not to obscure the present invention.

[0015] The present invention may be implemented by hardware, software, firmware, microcode, or any combination thereof. When implemented in software, firmware, or microcode, the elements of the present invention are the program code or code segments to perform the necessary tasks. A code segment may represent a procedure, a function, a subprogram, a program, a routine, a subroutine, a module, a software package, a class, or any combination of instructions, data structures, or program statements. A code segment may be coupled to another code segment or a hardware circuit by passing and/or receiving information, data, arguments, parameters, or memory contents. Information, arguments, parameters, data, etc. may be passed, forwarded, or transmitted via any suitable means including memory sharing, message passing, token passing, network transmission, etc.

[0016] The program or code segments may be stored in a processor readable medium or transmitted by a computer data signal embodied in a carrier wave, or a signal modulated by a carrier, over a transmission medium. The "processor readable medium" may include any medium that can store or transfer information. Examples of the processor readable medium include an electronic circuit, a semiconductor memory device, a read-only memory (ROM), a flash memory, an erasable ROM (EROM), a floppy diskette, a compact disk ROM (CD-ROM), an optical disk, a hard disk, a fiber optic medium, a radio frequency (RF) link, etc. The computer data signal may include any signal that can propagate over a transmission medium such as electronic network channels, optical fibers, air, electromagnetic, RF links, etc. The code segments may be downloaded via computer networks such as the Internet, Intranet, etc.

[0017] It is noted that the invention may be described as a process which is usually depicted as a flowchart, a flow diagram, a structure diagram, or a block diagram. Although a flowchart may describe the operations as a sequential process, many of the operations can be performed in parallel or concurrently. In addition, the order of the operations may be re-arranged. A process is terminated when its operations are completed. A process may correspond to a method, a function, a procedure, a subroutine, a subprogram, etc. When a process corresponds to a function, its termination corresponds to a return of the function to the calling function or the main function.

[0018] Figure 1 is an exemplary diagram illustrating a system 100 in which one embodiment of the invention can be practiced. The system 100 includes ring 110_{k-1}, L rings $110_{k,1}$ ... $110_{k,1}$ ...$110_{k,L}$, M rings $110_{k+1,1}$ ... $110_{k+1,m}$ .. . $110_{k+1,M}$.

[0019] The system 100 is a model of a hierarchical ring structure of the peer nodes connected to a network. The hierarchical ring structure is similar to a tree structure where there are a number of levels. For illustrative purposes, three levels k-1, k and k+1 are shown. However, the number of levels may be less than three or more than two. Also, a level may have any number of rings. Rings at level k-1 are referred to as upper rings which are at one hierarchical level higher than level k. Rings at level k are referred to as the current rings. Rings at level k+1 are referred to as lower rings which are at one hierarchical level lower than level k.

[0020] Each ring or group has a number of peers connected together. Here, the term "peer" refers to a node, a network device, or a computer that participates in a group activity in a network. A group activity may involve a number of peers that are linked together for a common objective. For example, users of shared programs may form into peers in an activity of shared programs. The group activity may be, for example, a business meeting, a multi-player game, a music sharing, etc. The term "connected," "linked," "connection," or "link" here may mean a physical connection and/or a logical or informational connection. When a line shows a connection or link between two peers, it means that one peer contains information of the other peer. The information includes at least an IP address of the peer.

[0021] Normally, a peer at an upper level contains information of a peer at a current level. It is contemplated that a peer at a lower level may also contain information of a peer at a current level. A peer in any ring is connected to any other peer via the hierarchical ring structure. This is possible because, in addition to being connected to peers within a ring, a peer may be may be connected to one or more peers at the upper level and one or more peers at the lower level.

[0022] A ring s at a level k has a number of peers. In each ring, all peers are connected together, meaning that each peer has information about any other peers within the same ring. In Figure 1, the notation for the number of peers for a ring s at level k is P(k,s). Ring 110_{k-1} has P(k-1,1) peers 120_{k-1}^{1} to $120_{k-1}^{P(k-1,1)}$. Also, in this illustrative example, level k-1 has one ring. At level k, there are L rings. At level k+1, there are M rings.

[0023] A peer P in a ring may be connected to one or more peers at a level immediately lower than its level. The ring at the upper level is the index ring of peers. In addition to caching information of other peers in the ring, each peer of an index ring also contains information of the lower level. The fan-out ratio, i.e. the number of lower level peers that each upper level index peer contains, can be selected according to some network or performance criteria to optimize or improve network performance. Similarly, the maximum number of peers MAX in a ring at a level may also be pre-determined. Alternatively, the fan-out ratio and/or the value of MAX may be selected dynamically as the ring structure changes depending on some dynamic behaviors such as traffic flow, quality of service, delays, fault-tolerance criteria, etc. Furthermore, the number of levels may be determined in advance statically or dynamically according to some dynamic

behaviors as discussed above. There are, however, at least two levels in the hierarchical ring structure.

[0024] A peer may connect to one or more peers at the lower level. Furthermore, more than one peer may connect to the same peer at the lower level. This kind of redundancy helps improve fault tolerance. For example, if peers P1 and P2 are connected to peer Q at the lower level and if peer P1 becomes faulty, then peer P2 still contains the information of peer Q. Therefore, any discovery or search process looking for information on peer Q can still be found via peer P2.

[0025] In the illustrative diagram shown in Figure 1, peer $120_{k-1}{}^{m}$ is connected to peers $120_{k,1}{}^{u}$ of ring $110_{k,1}$ and peer $120_{k,1}{}^{v}$ of ring $110_{k,1}$. In other words, peer $120_{k-1}{}^{m}$ contains information regarding peers $120_{k,1}{}^{u}$ and $120_{k,1}{}^{v}$ in addition to all peers within its own ring. Peer $120_{k,1}{}^{w}$, peer $120_{k,L}{}^{x}$ and peer $120_{k,L}{}^{y}$ are all connected to peer $120_{k+1,M}{}^{t}$. Therefore, the information on peer $120_{k+1,M}{}^{t}$ is contained in all three peers $120_{k,1}{}^{w}$, $120_{k,L}{}^{x}$ and $120_{k,L}{}^{y}$. Since the structure does not have a single point of failure, it is highly fault-tolerant and highly available. In addition, the structure is highly scalable. There is no limit on the number of levels in the structure. Levels or rings may be added or deleted as the network of peers grows or shrinks. Unlike pure P2P technology which may create a lot of network traffic, the hierarchical ring structure does not generate as much traffic structure because peer locating or discovery occurs within peers that are connected and not all peers in the network.

[0026] Figure 2 is an exemplary diagram illustrating a peer 200 in which one embodiment of the invention can be practiced. For clarity, the system of subscripts and superscripts are dropped. The peer 200 may represent any peer shown in Figure 1. The peer 200 includes a processor 110, a host bus 120, a memory control hub (MCH) 130, a system memory 140, an input/output control hub (ICH) 150, a mass storage device 170, input/output (I/O) devices 180_1 to 180_K, and a network device 185. A peer may include more or less elements than these elements.

[0027] The processor 110 represents a central processing unit of any type of architecture, such as embedded processors, micro-controllers, digital signal processors, superscalar computers, vector processors, single instruction multiple data (SIMD) computers, complex

instruction set computers (CISC), reduced instruction set computers (RISC), very long instruction word (VLIW), or hybrid architecture.

[0028] The host bus 120 provides interface signals to allow the processor 110 to communicate with other processors or devices, e.g., the MCH 130. The host bus 120 may support a uni-processor or multiprocessor configuration. The host bus 120 may be parallel, sequential, pipelined, asynchronous, synchronous, or any combination thereof.

[0029] The MCH 130 provides control and configuration of memory and input/output devices such as the system memory 140 and the ICH 150. The MCH 130 may be integrated into a chipset that integrates multiple functionalities such as the isolated execution mode, host-to-peripheral bus interface and memory control. For clarity, not all the peripheral buses are shown. It is contemplated that the system 100 may also include peripheral buses such as Peripheral Component Interconnect (PCI), accelerated graphics port (AGP), Industry Standard Architecture (ISA) bus, and Universal Serial Bus (USB), etc.

[0030] The system memory 140 stores system code and data. The system memory 140 is typically implemented with dynamic random access memory (DRAM) or static random access memory (SRAM). The system memory may include program code or code segments implementing one embodiment of the invention. The system memory includes a peer locator module 145, a peer interface module 146 and a registrar module 148. Any one of the peer locator module 145, the peer interface module 146 and the registrar module 148 may also be implemented by hardware, software, firmware, microcode, or any combination thereof. The system memory 140 may also include other programs or data which are not shown, such as an operating system.

[0031] The ICH 150 has a number of functionalities that are designed to support I/O functions. The ICH 150 may also be integrated into a chipset together or separate from the MCH 130 to perform I/O functions. The ICH 150 may include a number of interface and I/O functions such as PCI bus interface, processor interface, interrupt controller, direct memory access (DMA) controller, power management logic, timer, universal serial bus (USB) interface, mass storage interface, low pin count (LPC) interface, etc.

[0032] The mass storage device 170 stores archive information such as code, programs, files, data, applications, and operating systems. The mass storage device 170 may include compact disk (CD) ROM 172, floppy diskettes 174, hard drive 176, and any other magnetic or optic storage devices. The mass storage device 170 provides a mechanism to read machine-readable media. In particular, the mass storage device 170 contains a P2P cache subsystem 175 that is used to keep tract of peer information. The cache P2P subsystem 175 may use any one of the peer locator module 145, the peer interface module 146 and the registrar module 148, or it may contain its own processor and hardware and/or software modules having the same functionalities.

[0033] The I/O devices 180_1 to 180_K may include any I/O devices to perform I/O functions. Examples of I/O devices 180_1 to 180_K include controller for input devices (e.g., keyboard, mouse, trackball, pointing device), media card (e.g., audio, video, graphics) and any other peripheral controllers.

[0034] The network device 185 provides interface to a network 190. The network device 185 has the proper protocol and interface circuitry to allow connections to the network 190. The network 190 is any private or public network that supports P2P communication. The network 190 may be a local area network (LAN), a wide area network (WAN), the Internet, an intranet, or an extranet.

[0035] Figure 3 is an exemplary diagram illustrating the P2P subsystem 175 shown in Figure 2 according to one embodiment of the invention. The subsystem 175 may be implemented as a combination of hardware and software. The peer locator module 145, the peer interface module 146 and the registrar module 148 shown in Figure 2 may be used as part of the subsystem 175. The subsystem 175 includes a cache 310, a peer locator 320, a peer interface 330, and a registrar 340. It is contemplated that the subsystem 175 may contain more or less the components as shown.

[0036] The cache 310 is a mass storage device that can store information cached from the P2P communication. The information includes connectivity information such as the address information of the peers that are connected to the peer 120. This includes the information of the target peer if it is one of the peers that are connected to the peer 120. The cache 310 is used by a current peer 120 in a current ring at a current level to store

information of ring peers within the current ring. As described in Figure 1, the current ring is part of the hierarchical ring structure 100 of P2P nodes. The hierarchical ring structure has at least one of a lower level and a higher level with respect to the current level. The cache 310 stores a current level information 312, a lower level information 314 and a higher level information 316. The current level information 312 includes the information of the peers connected to the same ring as the current ring. The lower level information 314 includes the information of all the peers at the lower level that are connected to the current peer 120. The higher level information stores information of all peers at the higher level that are connected to the current peer 120. In one embodiment, the higher level information 316 is not cached.

[0037] The information stored in the cache 310 does not have a Time-To-Live (TTL) field. The cached list of peers is pinged at a present time interval to check if these peers are still alive or connected to the current peer 120. If these peers are dead, the associated peer information in the cache 310 is taken out of the cache 310. Valid or current information of new peers is then requested. Since peer information is updated periodically, no stale information is kept in the cache 310 for a long time.

[0038] The peer locator 320 locates a target peer in the cache 310 in response to a request. The request may be initiated by a user of the peer 120, or received by a requesting peer from the current ring or from a lower level or a higher level. As discussed above, the peer locator 320 may use the peer locator module 145 or may be implemented as a separate component using hardware, software or a combination of hardware and software. The peer locator 320 includes an information retriever 325 to retrieve the information of the target peer from the cache 310 if the target peer is located in the cache 310.

[0039] The peer interface 330 interfaces to others peers in the system. As shown in Figure 1, this peer may be a peer in the same ring, at a lower level and at a higher level. The peer interface 330 may also interface to peers in different rings at the same level, either lower level or higher level. Also as discussed above, the peer interface 330 may use the peer interface module 146 or may be implemented as a separate component using hardware, software or a combination of hardware and software. The peer interface 330 forwards the request to search the target peer to a peer at the lower level or a peer at the upper level when the target peer is not located in the cache 310. In one embodiment, the peer interface

330 forwards the request only to a peer at the upper level that is connected to the current peer 120. The peer interface 330 includes a lower interface 332 and an upper interface 336.

[0040] The lower interface 332 interfaces to peers at the lower level. These peers may include peers that are in different rings or in the same ring. The lower interface 332 forwards the request to at least one of these lower peers in ring {K...M} to search for the target peer when the target peer is not located in the cache 310. The lower interface also receives the request from one of the lower peers that is connected to the current peer 120 to search for the target peer. When this occurs, the lower interface 332 passes the request to the peer locator 320 which then carries out the task of locating the target peer.

[0041] The upper interface 336 interfaces to peers at the upper level that are connected to the current peer 120. These upper peers in ring {I...J} may be in the same ring or in different rings. The upper interface 336 forwards the request to at least one of these upper peers to search the target peer when the target peer is not located in the cache 310. The upper interface 336 also receives the request from one of these upper peers to search the target peer. When this occurs, the upper interface 336 passes the request to the peer locator 320 for carrying out the search task.

[0042] The registrar 340 processes registration of the current peer 120 to the upper peers that are connected to the current peer 120. The registrar also processes registration of the lower peers to the current peer 120. When the current peer 120 registers to an upper peer, it transmits its information including its address and other pertinent information so that the upper peer can store this information in its cache for peer locating or discovery upon request. Similarly, when a lower peer registers to the current peer 120, it sends its information including its address and other pertinent information to the current peer 120 so that the current peer can store this information in the cache 310 for peer locating or discovery upon request.

[0043] Figure 4 is an exemplary flowchart illustrating a process 400 for P2P communication using a hierarchical ring structure according to one embodiment of the invention.

[0044] Upon START, the process 400 receives a request or query from a requesting peer to locate or discover a target peer (Block 410). The requesting peer is normally a peer within the same ring or a lower peer having connection with the current peer. It is contemplated that the request may also come from an upper peer having connection with the current peer. The requesting peer has registered its information to the current peer. Then, the process 400 searches for the target peer in the cache (Block 420).

[0045] Next, the process 400 determines if the target peer is located in the cache (Block 430). If the target peer is not located in the cache, the process 400 forwards or escalates the request to an upper peer that is connected to the current peer (Block 440) and is then terminated. In one embodiment, the request is first forwarded or circulated to the peers in the same ring of the same hierarchy. If none of the peers in the same ring have the connectivity information in their respective caches, the request is sent to the upper peer. Note that in alternative embodiments, the request may be sent to the peers in the same ring either simultaneously or after sending the request to an upper peer. Also, the request may be sent to a lower peer simultaneously, before or after sending the request to the peers in the same ring and/or to an upper peer.

[0046] If the target peer is found in the cache, the process 400 retrieves the information on the target peer from the cache (Block 450). The information includes the address of the target peer. Thereafter, the process 400 returns the requested information to the requesting peer (Block 460) and is terminated.

[0047] While this invention has been described with reference to illustrative embodiments, this description is not intended to be construed in a limiting sense. Various modifications of the illustrative embodiments, as well as other embodiments of the invention, which are apparent to persons skilled in the art to which the invention pertains are deemed to lie within the spirit and scope of the invention.